Project Endings

Team meeting, August 29, 2016

Present: Stewart Arneil, Claire Carlin, Ewa Czaykowska-Higgins, Lisa Goddard, Martin Holmes, Tye Landels, Greg Newton, Joey Takada,

Although work on Graves, *MoEML* and *Mariage* is progressing simultaneously, Martin will use Graves as the first over-all test case for building an HTML-based eXist web application as it is the smallest of the group and is much less complicated. All of the projects, although built statically for reasons of simplicity and to aid in archiving, still require relatively sophisticated search capabilities, and the initial approach to this will be to use the eXist XML database, since we know it well.

1. The Graves Diaries

Tye and Joey made major progress over the summer.

* Tye did the clean up. There are still a couple of missing references Elizabeth will work on. Diagnostics show only 4 anomalies, all on one topic (“Estanco”).
* Joey wrote the Schematron and added to the diagnostics; he will follow up on a couple of problems.
* Joey is also working on the static build. P5 docs needed to be checked, but the encoding is already much improved, and the static build up and running; it looks nearly identical to the original site. It works better with JavaScript, and looks good; it works without JS, but of course doesn’t look very good. The project is ready to be packaged when the search is written. The site index on main page is a substantial improvement. The site was previously “uncrawlable” by harvesters; now a crawler will be able to find everything right away. We will test this using the Heritrix crawler.

When archived, HTML will be searched, rather than XML. There will be some work involved in configuring the indexing so that it will work intuitively.

2. *Mariage*

* Martin will work on *Mariage* after Graves to develop the search page and web app.
* Claire made less progress than planned in proofreading.
* Both have contacted colleagues in France to solicit MITACS applications. The MITACS website is entirely dependent on Flash, with the attendant problems (and risk of alienating DH colleagues). The September 20 deadline may be problematic for French students.
* Building of the *Mariage* web app will initially be done using the “tiny site” that is created as part of the static build, for testing purposes; the scale of this site should enable us to test the build process more rapidly.

3. MoEML

Every page that needs to exist has now been created as part of the static build.

Martin is beginning to build the web app, but will probably continue with *Mariage* first (search page and web app).

4. Moses

Ethan has done 250 hours; he is two-thirds of the way through the dictionary (there are some 3000 entries left). Colville tribes have signed one consent form; the other has been submitted to an elder for approval. The Colville tribes have ordered copies of substantial archival material from the University of Washington Suzzallo Library which will be shared with Ewa. Eventual links to this material would be a great contribution, though perhaps difficult to realize.

Ewa and Martin’s paper on using XML and TEI is having some impact among linguists, having been cited at least once. Summer Institute of Linguistics wrote to Ewa to ask for access to Moses database as a model for conversion to XML; they’re also interested in TEI but don’t understand the challenges of conversion, especially given the idiosyncrasy of Moses TEI. It would make more sense for them to work from different samples; Joey volunteered to prepare some samples for them. The Colville tribes are OK with representative examples being shared, as long as permission is obtained. Martin remarked that improving our TEI on all projects is a goal of Project Endings.

All four case studies

Images; the storage challenge. Stand alone eXist apps are the goal; they are completely portable. However, several of the projects have large archives of images which ought also to be stored as archival objects in the Library’s future replacement for ContentDM. What would the Library suggest? A paper could come out of this.

Once projects are working, perhaps work with Solar, but can it effectively index the data? First, try it on the static sites; the search page points at Solr rather than eXist. This might well work; experimentation is needed. Lisa noted that the goal should be something that scales over time and is automatable. Our plans, developed with Ben at the Library, involve creating and testing (on our own Solr instance) Solr configuration files and settings for each of the projects; when the projects are ready for archiving, these Solr configs can be passed on to Ben, who can implement them on the Library’s system. We expect to run the eXist web applications with their built-in custom search functions as the main “live” front end, but work on making the Solr search of the same data as good as the eXist search; if that can be achieved, the projects can be converted to static archives served by a regular web server, and we can move away from eXist.

Joey asked what type of metadata would be used for objects. Martin noted that it should be placed in the header of HTML pages; the best current model is the *Colonial Despatches*, along with some of what already exists in *MoEML* (keywords, etc.). Joey suggested embedding Dublin Core metadata in the TEI headers, in the <xenoData> element, but Martin see little value in this approach.

Lisa: Dublin Core mapped. Should Joey do Dublin Core for Graves? He will talk to Dean Seeman, head of cataloguing, who is currently working on DC data models for the Library’s developing systems. Lisa noted the need to align data models.Martin added that HTML is what the world will see.

Library

* The Library continues to build infrastructure. They are offering to faculty the opportunity to put digital objects on line through a web form matched to Dublin Core fields. The result will be searchable and findable using Spotlight, which is also an exhibit builder. These systems don’t load XML in any useful way.
* Joey: what about images? Lisa: this system generates tiffs, jpegs (the access image), thumbnails.
* The Library should work to avoid duplication with HCMC.
* Martin: there is a distinction to be made between archived images and image material crucial to the site that doesn’t need to be formally archived; there is a difference between an archival image of an early modern print page and a photograph from a project team party. All images need to be part of the web application, but only a subset will need formal archiving in the Library system.

HCMC move:

Construction is supposed to be finished at the end of November. Lisa thinks January is more likely. The HCMC will be down for at least two-three days – or longer…